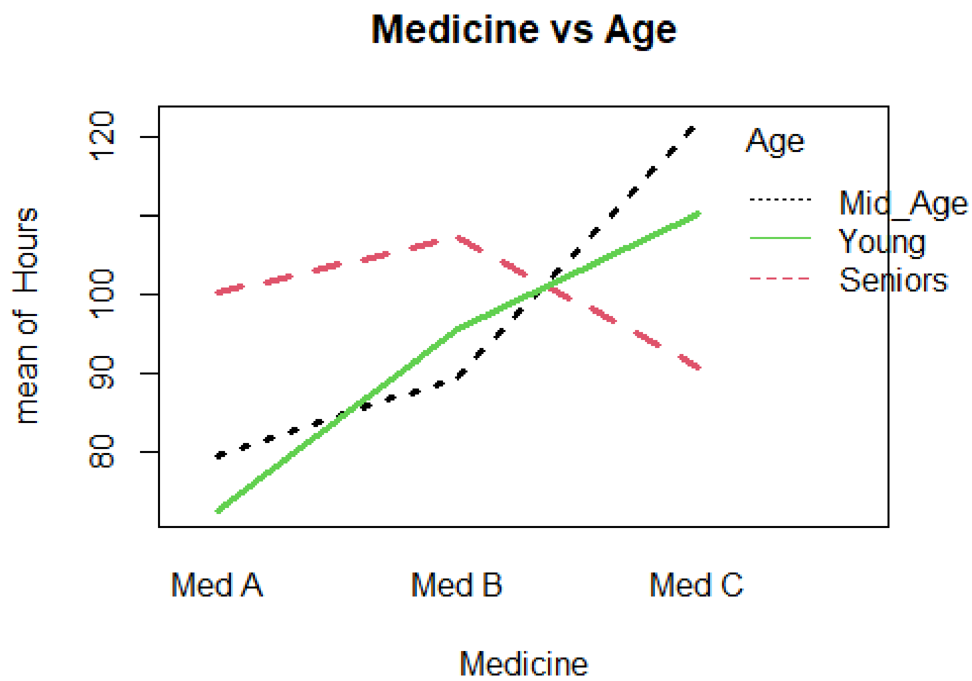
Test 3

Prasanth Chowdary Yanamandala

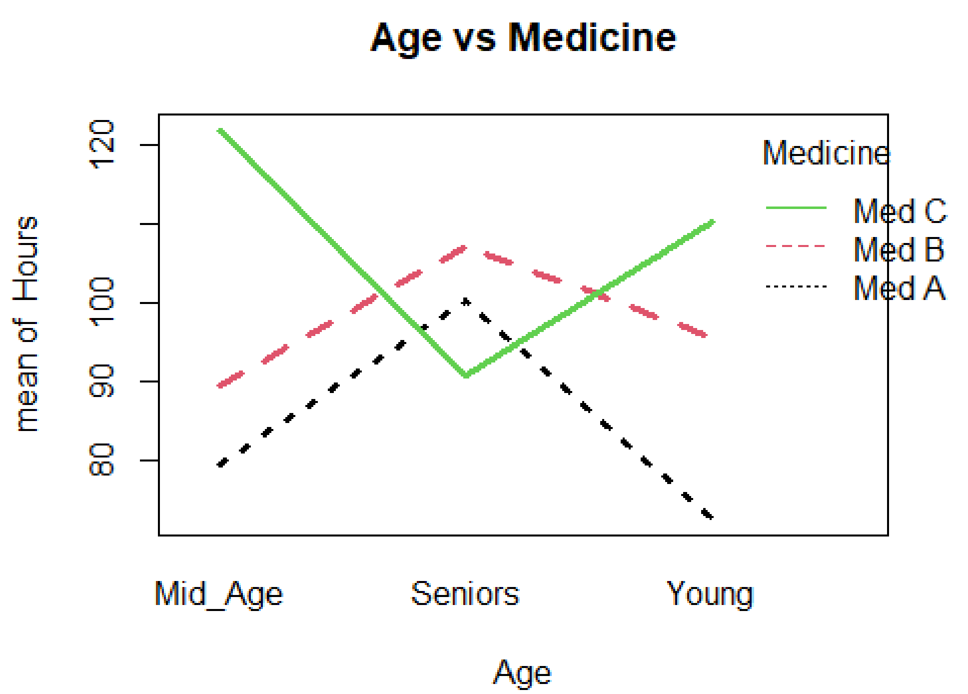
2023-12-06

Problem 1 (Set-1)
ANOVA 2-Factor
Three anti-bacteria creams were used on three age groups. The number of hours before the medicines started to show a noticeable effect are recorded in the table. Assume α = 0.05
a. Run this as an ANOVA 2-factor R program.
b. Also draw the interaction graph to show the interaction between the two factors.
c. Answer the online test questions.

```
#Clear the environment
rm(list = ls())

#  set the working directory to dir1
dir1 <- getwd()
setwd(dir1)

#load readxl library
library(readxl)
library(e1071)

#Read excel sheet for Problem 1
Data1 <- read_excel("F23-6359-Test-3.xlsx",sheet="Set-1")

#Run Anova 2 factor
attach(Data1)
A1<-aov(Hours ~ Medicine + Age + Medicine:Age)
summary(A1)

##              Df Sum Sq Mean Sq F value  Pr(>F)
## Medicine      2   8311    4156   6.082 0.00346 **
## Age           2    674     337   0.494 0.61230
## Medicine:Age  4  10042    2511   3.674 0.00843 **
## Residuals    81  55350     683
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

#interaction plot
interaction.plot (Medicine, Age, Hours, lwd = 3,  col=1:3,main="Medicine vs
Age")
```




```
interaction.plot (Age, Medicine, Hours, lwd = 3,  col=1:3,main="Age vs
Medicine")
```



```r
detach(Data1)
```

Null Hypothesis: -The population means of all ages are equal. -The population means of all medicines are equal. -There is no interaction between age and medicine.

Conclusion: -age's overall effect is the same. -medicine's effect make a difference. -no consistency of age's effect across the medicines.

2. the P-value for Age is 0.61230
3. the P-value for Medicine is 0.00346

4. the F-stat for medicine is 6.082

Problem 2 (Set-2)
Two sample t-test
Two teaching methods are being researched in a big school district of Texas. Random samples were taken to see if the two methods are the same or different.

a. Do a variance test to see if the two variances are equal.

b. Do the appropriate t-test at $\alpha = 5\%$.

c. Answer the online test questions.

```r
#Clear the environment
rm(list = ls())

#Read excel sheet for Problem 2
Data2 <- read_excel("F23-6359-Test-3.xlsx",sheet="Set-2")

#Variance Test
attach(Data2)
var.test(Scores ~ Method)

##
##  F test to compare two variances
##
## data:  Scores by Method
## F = 2.7273, num df = 99, denom df = 95, p-value = 1.492e-06
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  1.82584 4.06709
## sample estimates:
## ratio of variances
##           2.727323

#t-test
ttest <- t.test(Scores ~ Method, var.equal=FALSE, alternative="two.sided")
ttest
```

```
##
##  Welch Two Sample t-test
##
## data:  Scores by Method
## t = -1.6038, df = 164.12, p-value = 0.1107
## alternative hypothesis: true difference in means between group Method_A
and group Method_B is not equal to 0
## 95 percent confidence interval:
##  -6.5567858  0.6792858
## sample estimates:
## mean in group Method_A mean in group Method_B
##               73.03000               75.96875
```

```
detach(Data2)
```

Null Hypothesis: -Variances are equal for f test -Difference in means of two methods is 0 for t-test Alternative Hypothesis -Variances are not equal for f test -Difference in means of two methods is not 0 for t-test

Conclusion -P value is less than alpha for F test so we reject Null Hypothesis -P value is more than alpha for T test so we fail to reject Null Hypothesis

8.p-value from the Variance test is 0.000001492 11.the p-value for the t-test is 0.1107 12.what is the t-statistics?: -1.6038 ——————————————————————————————————————– Problem 3 (Set-3)
Logistics Regression
A College in Texas is trying to figure out how Interships are offered. It has collected a sample of 345 students.
The GPA (out of 10), Interview Score (Out of 70), and Aptitude Test score (out of 50) are given in Set-3.
Based on this data, answer the online test questions using your R code.

```
#Clear the environment
rm(list = ls())

#Read excel sheet for Problem 3
Data3 <- read_excel("F23-6359-Test-3.xlsx",sheet="Set-3")

library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
#logistic regression
attach(Data3)
Int<-glm(Internship ~ GPA + Interview + `Apt test`, family="binomial")
summary(Int)
```

```
##
## Call:
## glm(formula = Internship ~ GPA + Interview + `Apt test`, family =
"binomial")
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -8.53909    1.77951  -4.799  1.6e-06 ***
## GPA          0.24928    0.10759   2.317  0.02050 *
## Interview    0.04607    0.02137   2.156  0.03105 *
## `Apt test`   0.08615    0.02792   3.086  0.00203 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 455.05  on 344  degrees of freedom
## Residual deviance: 431.42  on 341  degrees of freedom
## AIC: 439.42
##
## Number of Fisher Scoring iterations: 4
```

```r
coef(Int)   # Maximum likelihood estimates of the parameters
```

```
## (Intercept)         GPA   Interview  `Apt test`
## -8.53908981  0.24928006  0.04607372  0.08614607
```

```r
RegOut<-c(coef(Int))
RegOut
```

```
## (Intercept)         GPA   Interview  `Apt test`
## -8.53908981  0.24928006  0.04607372  0.08614607
```

```r
#probability of getting Internship
Odd1<-exp(RegOut[1]+RegOut[2]*8+RegOut[3]*70+RegOut[4]*49)
p1<-Odd1/(1+Odd1)
p1
```

```
## (Intercept)
##   0.7112513
```

```r
#odds of getting Internship
Odd49<-exp(RegOut[1]+RegOut[2]+RegOut[3]+RegOut[4]*49)
Odd39<-exp(RegOut[1]+RegOut[2]+RegOut[3]+RegOut[4]*39)
Odd2<-Odd49/Odd39
Odd2
```

```
## (Intercept)
##    2.366615
```

```
Odd3<-exp(RegOut[1]+RegOut[2]*6+RegOut[3]*38+RegOut[4]*36)
Odd3
```

```
## (Intercept)
##   0.1117658
```

```
detach(Data3)
```

14.Intercept calculated by R: -8.53908981 15.coefficient of GPA as calculated by R: 0.24928006 16.coefficient of Interview as calculated by R: 0.04607372 17.what is the probability of getting Internship for a student whose GPA is 8, Interview score 70 and Aptitude score is 49?: 0.7112513 18.what are the odds of getting internship of someone who scored 49 on the Aptitude test vs some who who scored 39?: 2.366615 19.what are the Odds of a student getting Internship whose GPA is 6, Interview is 38 and Aptitude test score is 36?: 0.1117658 ——————————————————————————————– Problem 4 (Set-4) Log Transformation You've picked up a bunch of rocks from a rocky beach and want to estimate the weight of all the rocks at the beach with a Confidence level of 93.47%.

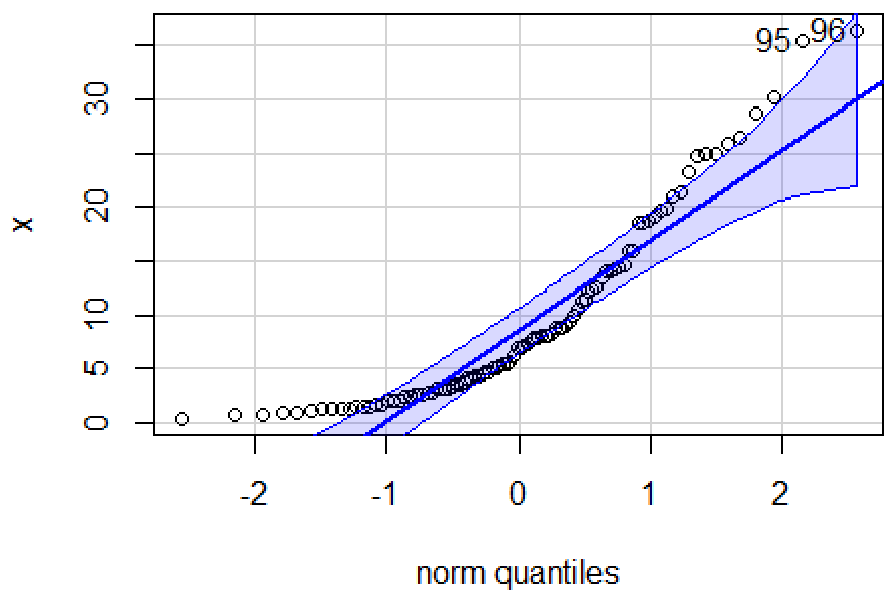
a. Plot the qqline and boxplot of the data. Also get the skewness. What is your conclusion about the distribution being normal? b. Do a log transformation (base e) and perform the steps in a. What's your conclusion?

c. Use Log transformed data to answer the online test questions.

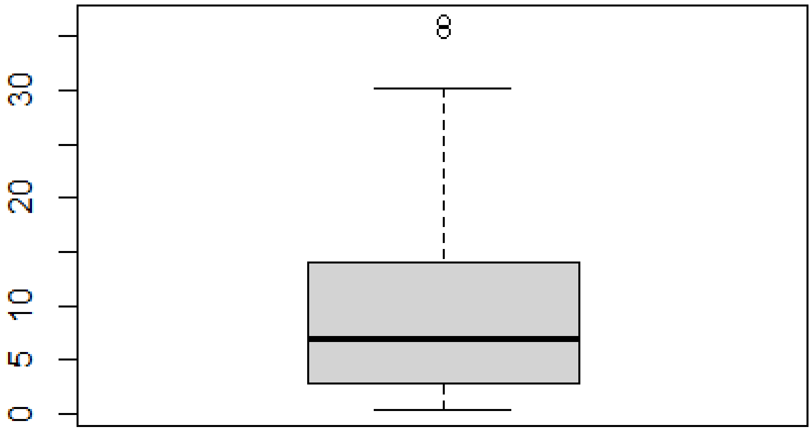
```
#Clear the environment
rm(list = ls())

#Read excel sheet for Problem 4
Data4 <- read_excel("F23-6359-Test-3.xlsx",sheet="Set-4")

attach(Data4)
list(Data4)
```

```
## [[1]]
## # A tibble: 96 × 1
##    Weight
##     <dbl>
##  1   0.42
##  2   0.86
##  3   0.88
##  4   1.11
##  5   1.34
##  6   1.38
##  7   1.42
##  8   1.47
##  9   1.63
## 10   1.73
## # ℹ 86 more rows
```

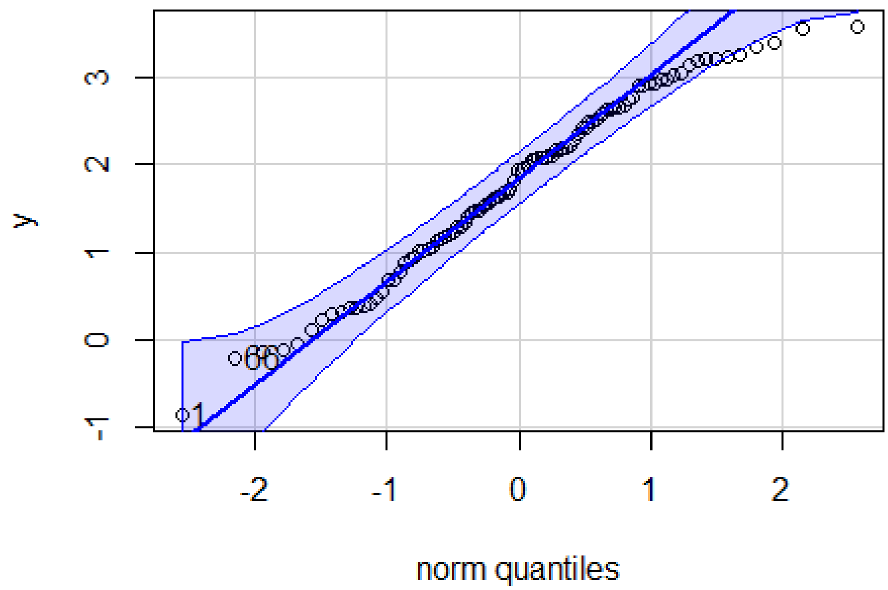
```r
#Log transformation
#before
library(ggplot2)
library(car)

## Loading required package: carData

##
## Attaching package: 'car'

## The following object is masked from 'package:dplyr':
##
##     recode

x<-Weight
#QQLine, boxplot and skewness
qqPlot(x)
```



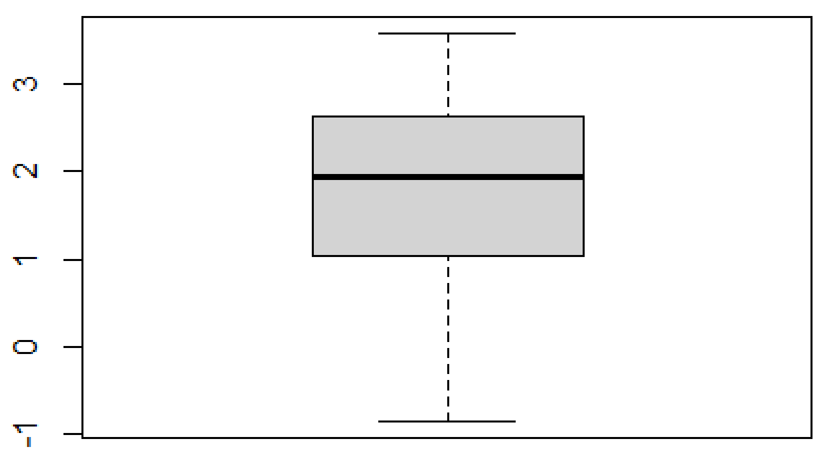

```r
## [1] 96 95

boxplot(x)
```




```
skewness(x)

## [1] 1.214707

#after
y<-log(Weight)
#QQline, boxplot and skewness
qqPlot(y)
```




```
## [1]  1 66

boxplot(y)
```



```
skewness(y)

## [1] -0.2823303

#mean of log data
m<-mean(y)
m

## [1] 1.79195

#skewness of log data
s<-skewness(y)
s

## [1] -0.2823303

#standard Deviation of log data
std<-sd(y)
std

## [1] 1.026756

#standard error of log data
stde<-std/sqrt(length(y))
stde

## [1] 0.1047928

#LCL before rev.trans using t.test
t.test(y, conf.level = 0.9347)

##
##  One Sample t-test
##
## data:  y
## t = 17.1, df = 95, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 93.47 percent confidence interval:
##  1.596535 1.987366
## sample estimates:
## mean of x
##   1.79195

t_score<-qt(p=0.0653/2,df=95,lower.tail = FALSE)
t_score

## [1] 1.864775

LCL.log<-m-(t_score*stde)
LCL.log

## [1] 1.596535
```

```r
#LCL after rev.trans using t.test
z<-exp(y)
t.test(z, conf.level = 0.9347)
```

```
##
##  One Sample t-test
##
## data:  z
## t = 10.84, df = 95, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 93.47 percent confidence interval:
##   7.792313 11.030187
## sample estimates:
## mean of x
##   9.41125
```

```r
t_score<-qt(p=0.0653/2,df=95,lower.tail = FALSE)
t_score
```

```
## [1] 1.864775
```

```r
LCL<-mean(z)-(t_score*(sd(z)/sqrt(length(z))))
LCL
```

```
## [1] 7.792313
```

```r
detach(Data4)
```

Before transformation the data was heavily skewed and not normally distributed with the data points scattered away from the normal line in qqplot and mean weighted on the lower side in box plot. After transformation the skewness reduced bringing the value closer to 0 and points gathered closer to the normal line in qqplot with the box plot situated in the middle instead of being weighted down.

20.what is the mean of the log-transformed data?: 1.79195 21.Skewness of the sample (after log transformation)?: -0.2823303 22.calculate the standard Deviation after log transformation: 1.026756 23.what is the standard error of the transformed data?: 0.1047928 24.lower limit of Confidence Interval for a Confidence level of 93.47% (before reverse transformation)?: 1.596535 25.what is LCL after reverse transformation?: 7.792313 ———————————————————————————– Problem5 (Set-5) ChiSq Tests A financial institution wants to evaluate it's policy of employing students based on the college reputation (high vs low tution).
Since employees' incomes also depends on the commission they earned, it took a sample of 866 employees to study if their college-type matters.
The table in Set 5 gives a breakdown of the data collected. Solve this problem using R and answer the online test questions.

```r
#Clear the environment
rm(list = ls())
```

```r
#Read excel sheet for Problem 4
Data5 <- read_excel("F23-6359-Test-3.xlsx",sheet="Set-5")
```

```
## New names:
## • `` -> `...1`
```

```r
x<- matrix(c(133,104,54,115,115,63,100,107,75), nrow=3)
View(x)
#Chi-sq test
chisq.test(x)
```

```
##
##  Pearson's Chi-squared test
##
## data:  x
## X-squared = 8.5879, df = 4, p-value = 0.07227
```

```r
qchisq(0.05,4,lower.tail = FALSE)
```

```
## [1] 9.487729
```

26.what is the P-Value?:0.07227 28.what are the degrees of Freedom?: 4 29.what is the ChiSq Critical? Assume Alpha = 5%.: 9.487729